Программа кластеризации векторных данных «datatree»

Назначение

Программа предназначена для представления множества многомерных векторов в виде дерева таким образом, что ближайшие вектора разделены наименьшим количеством узлов и длины ветвей представляют меру различия векторов (в данном случае эвклидово расстояние, но возможны и другие варианты).

В качестве кластеризуемого множества может выступать любое множество данных любой размерности. Программа была создана для кластеризации множества конформаций пептидов, полученного в ходе конформационного поиска.

Алгоритм

Программа выбирает два ближайших вектора и несколько (может быть ни одного) векторов по описанному далее алгоритму, и исключает их из кластеризуемого множества, заменяя их на их общий корень — их среднее арифметическое. Исключённые точки становятся дочерними узлами по-отношению к новому. Так повторяется до полного соединения всех векторов.

Программа использует два алгоритма выбора дополнительных векторов для объединения в один кластер. Первый — выбирает все векторы, включая два первых, такие что их расстояние ρ от их среднего были не более αρ, где α — коэффициент, задаваемый пользователем и $\alpha>1.0$. Второй алгоритм выбирает все векторы, формирующими связный граф, построенный так: граф включает первые две ближайшие точки с расстоянием ρ и любые две точки, расстоянием не более αρ, где α — коэффициент, задаваемый пользователем, так что $\alpha>1.0$ считаются связанными.

Входные данные

Во входной поток подаётся поток чисел. Размерность и количество векторов можно задать либо двумя первыми числами, либо параметрами -dim n -num N. Коэффициент α задаётся параметром -alpha ALPHA. Можно использовать параметр -input FILE для ввода из файла. Алгоритм кластеризации задаётся параметром -alg ALGNAME, где ALGNAME может быть либо spherical (для первого алгоритма) либо continuous (для второго).

Выходные данные

Программа печатает в выходной поток дерево в формате newick (*.ph), которое может быть просмотрено в программе treeviewx, njplot или другой подобной, рекомендуется использовать программу-просмотрщик с возможностью отображать длины ветвей. Можно использовать параметр -output FILE для вывода в файл.